Statistical Mechanics, Neural Networks, and Artificial Intelligence: *Using Powerful Brain Strategies to Improve AI*

Chapter 11 (DRAFT): Free Energy

Alianna J. Maren
Themesis, Inc.

Draft: 2024-01-15

11.1 Introduction and Overview

Many of us know of major universal laws, such as the *law of gravity*. Equally important, and governing how our universe works, is the *law of free energy minimization*.

Free energy is important; the reason that we may not be as familiar with it as we are with gravity is that it is the free energy of a system reflects a balance or interplay between two forces.

This is not the first time that we've encountered a universal law that reflects a balance between two forces. When the earth or any of our other planets are in a stable orbit around a sun, or when a spacecraft reaches orbital velocity and settles into a stable orbit around the earth, both the planets and the spacecraft are obeying a balance between two forces.

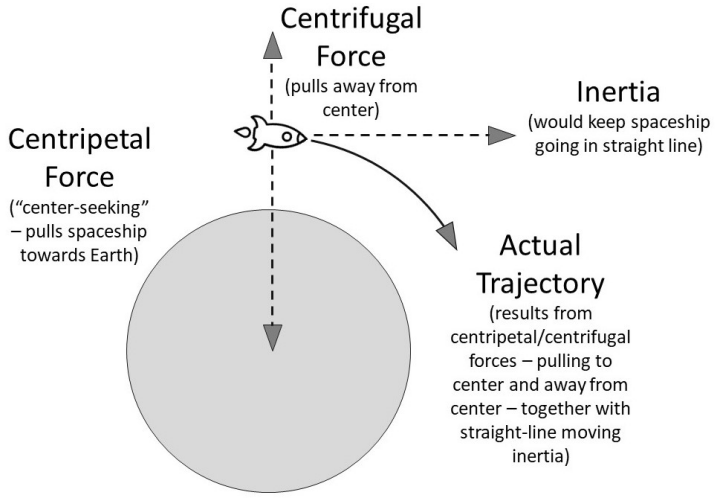

Figure 11.1: A spaceship stays in a stable orbit due to the balance between inward/outward forces (centripetal and centrifugal) together with a linear directional force (inertia) that moves the spaceship in a straight line tangentially away from the center of gravity.

The first of these is *centripetal force*, which is the gravity of the sun pulling the planets towards itself, or the gravity of earth pulling a spacecraft towards itself. The second force is the inertia of each moving object (the planet or the spacecraft). Newton described this as one of the *Laws of Motion*, saying that a body at rest will stay at rest, and a body in motion (in the absence of

some other force) will continue in motion, in the same direction and with the same constant velocity.

When we say that a spacecraft has "achieved orbital velocity," what we're saying is that the velocity of spacecraft (moving tangentially to an orbit around the earth) is sufficient to match the centripetal (gravitational) force of the earth pulling the spacecraft towards itself. Because the two forces are balanced, we have a stable state: the spacecraft continues in its orbit, without needing further acceleration (and expenditure of fuel).

Similarly, the *law of free energy minimization* provides stability, balance, and order in our universe. When the *free energy* of a system is minimized, we have a stable state, just as we have a stable orbit when gravity and motion are in balance.

The two component factors of free energy are *enthalpy* and *entropy*. We discussed *entropy* in the previous chapter, and will introduce the role of *enthalpy* here.

11.2 The free energy equation

The notion of free energy, which underlies a great deal of thermodynamics, is also important in neural networks, particularly those used in deep learning. Early neural networks, such as the Hopfield network and the Boltmann machine, relied on a free energy minimization approach.

We addressed these two neural networks, doing a contrast-and-compare of both their energy equations and their architectures, in a prior chapter. Now, we look at them again, this time probing deeper into how they are each governed by a free energy equation.

Even if we ignored these two networks (the Hopfield and the simple Boltzmann machine), we can't ignore the restricted Boltzmann machine (RBM), which is the heart and soul of deep learning. The RBM is a ***simplification*** (a *restriction*) on the simple Boltzmann machine.

Thus, the best way in which we can understandi the RBM is to understand the simple Boltzmann machine and its predecessor, the Hopfield neural network. This will give us a solid basis for understanding RBMs. Further, we'll have put down a strong foundation for understanding deep neural networks, which are created by stacking RBMs.

This chapter gives an altogether too-brief overview of free energy. Our primary goal, for now, is simply to ***recognize a free energy equation***

when we see one.

A secondary goal - which may be just as important for us practically - is to understand the *notation* used by various authors when they write the free energy equation.

Very often, we can understand the mathematics and derivations very well, but will get tripped up when we compare works by different authors (or even the same authors, over time), when they write about the same subject - but use different notation. This can cause a great deal of confusion.

Thus, we give particularattention to notation, as well as the actual formulation of the free energy equation.

As we'll observe, this is not as simple as it seems.

One challenge is ***that the free energy equation shows up in two remarkably different forms***. It's a lot like recognizing that a specific caterpillar corresponds with a certain specific butterfly. They look very different, but they are two expressions of the same creature.

A second challenge is that there is a ***range of notation used for free energy***. As we observed in the previous chapter, the notation for entropy could commonly be either an S or an H. For free energy, the notation can be F (for free energy, not surprisingly), or H or A (for Helmholtz free energy), or G (for Gibbs free energy).

These latter distinctions (Helmholtz vs. Gibbs) make a great deal of difference in the world of physical chemistry, where pressure and volume come into play. However, when we use free energy in the world of neural networks or machine learning (including variational methods), we are dealing with a *reduced* free energy, for which all the terms involving units of energy, temperature, or numbers of units in the system have been divided out. This leaves us with an equation that has no relation with the real, physical world. It is an abstraction; an *ideal*.

In the realm of neural networks and machine learning, using a *reduced* equation works well. Concepts such as the pressure or the volume of a system are not relevant. As a result, terms involving changes in those (pressure and/or temperature) variables drop out of the (reduced) free energy equation.

Occasionally, we (as readers) may come across authors who refer to Helmholtz or Gibbs free energies. For our work (in neural networks and machine learning), we can ignore those distinctions, and treat the Helmholtz free energy and the Gibbs free energy as the same thing; they are all the same "free energy" for our purposes.

Further, once the notion of free energy is fairly well understood, there are

a number of specific models that are common and well-known to physicists. These include the Ising model, together with its variants. One might read, for example, about a Bethé-Peierls or a mean-field model. We will ignore these variants in this treatment, and concentrate on the basics.

It is possible to derive the second form of the free energy equation (with a little calculus and elbow grease) from the first. For now, we ask you to take on faith that the two following equations mean the same thing.

The first formulation gives the free energy in terms of the energy of each unit, as encapsulated in the partition function.

The Free Energy - first version (as logarithm of the partition function:

$$F = -k_\beta T \ln(Z).$$

The most common way in which we see the free energy introduced in a paper uses the expression just given, so that the free energy involves the partition function.

Reading the First Version of the Free Energy Equation:

The free energy is the negative of a constant (Boltzmann's constant times temperature) times the natural logarithm of the partition function, Z.

There is an entirely different way of expressing free energy; as the difference between the enthalpy (or chemical potential, or ability to do work) minus the temperature times the entropy.

This is a crucial equation as we start using (free) energy minimization methods in machine learning. It means that the equilibrium, or minimal free energy state, is reached as a balance between getting the lowest possible energy (enthalpy) values while still maximizing entropy. This is a trade-off.

Eqn. 11.1 gives the second form of the free energy equation as

The Free Energy - second version (as the difference between enthalpy and temperature times entropy):

$$F = U - TS, \tag{11.1}$$

where U is the chemical potential, T is temperature and S is the entropy.

Reading the Second Version of the Free Energy Equation:

The free energy the difference between the chemical potential, U, and the temperature T times the entropy, S.

U, as mentioned previously, is the chemical potential, often defined as the ability of the system to absorb or release energy during chemical reactions. It can include a number of factors, most significantly (for our purposes) the enthalpy, which is the energy associated with each unit. Since the "other factors" do not come into play in machine learning, various authors may use either the term chemical potential or enthalpy. They may use U or E or H to express these terms, although U is generally reserved for the chemical potential, while E or H more commonly refer to the enthalpy.

As mentioned earlier, various letters are used for different terms, depending on the author's whim and provenance, in the free energy equation. Thus, we could see the Eqn. 11.1 show up as $G = H - TS$ or even $A = U - TH$, if the author wanted to be particularly confusing and substitute H (the information-theory way of expressing entropy) for S (the physical chemist's way of expressing entropy).

While the authors will usually define their terms, they occasionally leave interpretation up to their reader. Then, we have to infer what they mean from context.

As we mentioned previously, when we use free energy for neural networks or machine learning (including variational inference), we create a *reduced free energy equation*. We do this by dividing-through by any and all terms that have dimensions of energy and temperature, as well as the total number of units in the system.(This latter step is because free energy, enthalpy, and entropy are *extensive* properties - that is, they depend on the total number of

units in the system. So, for example, the free energy (the ability to do work) is greater in a steam engine than it is in a wood-burning stove, and is greater in a wood-burning stove than it is in a candle flame. By dividing through by the *number of available units*, we put all these systems (steam engine, stove, and candle flame) on the same level. This lets us focus our attention on the nature of the equation itself.

In this light, we will *reduce* our free energy equation and all the terms in it.

This would normally give us an equation that we would write in the following manner.

The *Reduced* Free Energy - we have divided-through by all terms involving units of energy or temperature, as well as the total numbers of units.

$$\bar{F} = \bar{U} - \bar{S}, \tag{11.2}$$

where $\bar{F}$ is the *reduced* free energyl, $\bar{U}$ is the *reduced* chemical potential, $\bar{S}$ is the *reduced* entropy, and we have divided through by the temperature T.

Now, we take one more step with simplifying our notation, and instead of putting the "bar" over our terms, we will use the original terms F, U, and S, with the understanding that we are now referring to *reduced* terms - just with the simplified ("no-bar") notation. This is the way in which we will usually see the free energy equation in neural networks and variational inference papers. This gives us the following (simplified) equation.

The *Reduced* Free Energy - we have divided-through by all terms involving units of energy or temperature, as well as the total numbers of units, but are no longer using the "bar" over the terms. The *reduced* nature is understood from here on.

$$F = U - S, \tag{11.3}$$

where F is now the *reduced* free energyl, U is now the *reduced* chemical potential, and S is now the *reduced* entropy. The notion of temperature T is no longer directly important in this equation; the influence of temperature is absorbed into coefficients in the enthalpy term(s).

Now, we take one more step in the world of notation - we replace U with H, and could (just as equivalently) use E. (Both are in common use in the neural networks and machine learning literature.)

This means that we write the previous equation in the following form

The *Reduced* Free Energy - we have replaced U with E.

$$F = H - S, \tag{11.4}$$

where F is the *reduced* free energy, H is the *reduced* enthalpy (we are now using that term instead of chemical potential), and S is the *reduced* entropy.

11.3 Sometimes Very Different Notation

Occasionally, we come across an author who uses ***all*** the possible notational variants, and sometimes even invents new ones. An example is this equation in Friston's 2013 paper, "Life as We Know It" [1].

In this Eqn. 2.7 by Friston, we see that in the last line, he is using F to refer to the free energy of the system. (He is using it as a function of three variables, s, a, and λ, which are not important in this discussion.)

Friston then uses G (which he refers to as "Gibbs free energy," which - as we've discussed - is not a necessary distinction) for the enthalpy term, and then uses H (the information-theory notation) for entropy.

The only way to deal with the notational conundrums presented by authors such as this is diligent comparison of the actual equation specifics with their known references. For example, we read that Friston identifies his last term as the "entropy of the variational density." We can verify this when we look at how that equation is actually expresed, later in his works.

Karl Friston (2013); "Life as We Know It."

Lemma 2.1 Free energy. *For any Gibbs energy* $G(\psi, s, a, \lambda) = -\ln p(\psi, s, a, \lambda)$*, there is a free energy* $F(s, a, \lambda)$ *that describes the flow of internal and active states*:

$$\left.\begin{aligned} f_\lambda(s,a,\lambda) &= -(\Gamma + R)\cdot\nabla_\lambda F, \\ f_a(s,a,\lambda) &= -(\Gamma + R)\cdot\nabla_a F \end{aligned}\right\} \quad (2.7)$$

$$\text{and} \quad \left.\begin{aligned} F(s,a,\lambda) &= -\int_\psi q(\psi|\lambda)\ln\frac{p(\psi,s,a,\lambda|m)}{q(\psi|\lambda)}\,d\psi \\ &= E_q[G(\psi,s,a,\lambda)] - H[q(\psi|\mu)]. \end{aligned}\right\}$$

Here, free energy is a functional of an arbitrary (variational) density $q(\psi|\lambda)$ *that is parametrized by internal states. The last equality just shows that free energy can be expressed as the expected Gibbs energy minus the entropy of the variational density.*

Figure 11.2: Equation taken from Karl Friston's 2013 paper, "Life as We Know It."

11.4 Enthalpy as a Linear Combination of Two Terms

We often see that the enthalpy term is expressed as the sum of two terms, as shown in the following equation.

The enthalpy of a system:

$$H = E = \langle e_i \rangle + \langle e_{ij} \rangle.$$

The enthalpy of the system (which we previously identified as chemical potential) is often given as the sum of two terms; one expressing the expected energy associated with each individual unit, and the other expressing the energy associated with pairwise interactions between the units.

Reading the Enthalpy Term:

The enthalpy (which we previously considered as the chemical potential) is the sum of the expected energy per unit, e_i, together with the pairwise interaction between units, $e_{i,j}$ (Note that we are changing the meaning of the subscript j here; it now refers to another unit in the system, and not to a distinct microstate. This is to make it easier to read the next example, which quotes from one of John Hopfield's papers.

This formulation draws from the well-known statistical mechanics model known as the *Ising equation*. It is *very* common in energy-based neural networks.

As an example, we see how it is used in the Hopfield neural network. (The same thought-process applies to the Boltzmann machine, the restricted Boltzmann machine, and all its derivatives and descendents.)

John Hopfield's 1982 work presented a new idea in neural networks, using an energy function to describe the state of the network.

We looked at this equation in a previous chapter. At that point, though, we were just using this energy equation to help us understand what the *structure* of the neural network had to be. We were constructing an isomorphism between the structure as implicitly expressed in the equation, and the structure that we would create and train with different patterns.

Now, our purpose in looking at this equation is different. We're noting the correlation between an expression that defines how a neural network works and its motivating source in statistical mechanics.

We've already identified (in a previous chapter) that the two units, V_i and V_j, correspond to nodes in a neural network. We previously identified that these were defined to be *bistate* units; meaning that they could be in one of two states, and we gave those two states the values of "0" or "1."

Now, though, we're paying attention to the other term in this equation; $T_{i,j}$, which refers to the interaction energy between the two units V_i and V_j.

If we're familiar with statistical mechanics, then as soon as we see this equation, we know that we're dealing with a free energy approach. We further know that we're focusing on the *interaction energy* (more properly, the *interaction enthalpy*) of the system.

Extract from J. Hopfield (1982), Neural networks and physical systems with emergent collective computational abilities

Studies of the collective behaviors of the model

The model has stable limit points. Consider the special case $T_{ij} = T_{ji}$, and define

$$E = -\frac{1}{2}\sum_{i \neq j}\sum T_{ij} V_i V_j \ . \qquad [7]$$

ΔE due to ΔV_i is given by

$$\Delta E = -\Delta V_i \sum_{j \neq i'} T_{ij} V_j \ . \qquad [8]$$

Thus, the algorithm for altering V_i causes E to be a monotonically decreasing function. State changes will continue until a least (local) E is reached. This case is isomorphic with an Ising model. T_{ij} provides the role of the exchange coupling, and there is also an external local field at each site. When T_{ij} is symmetric but has a random character (the spin glass) there are known to be many (locally) stable states (29).

Figure 11.3: Extract from John Hopfield's 1982 paper on "emergent collective computation."

Example 11.1.
Suppose that you were to read John Hopfield's original paper, introducing what we now call the Hopfield neural network [2]. Figure 11.3 gives an extract from this paper. Reading this, we would note the equation

$$E = -\frac{1}{2}\sum_{i\neq j}\sum_{j} T_{i,j}V_iV_j$$

Hopfield's next equation gives us an expression for ΔE, which tells us how the energy changes over time. He specifically says that ΔE is a "monotonically decreasing function," meaning that the energy of the system is always either decreasing or holding steady; it never increases. This tells us that we're dealing with a free energy minimization approach.

When he further says that "the case is isomorphic with an Ising model," he's saying that we're very similar in our method to one of the classic models in statistical mechanics, where units can be either "on" or "off" (as with many neural networks), and that there is a prescribed energy-of-activation for the "on" units, and there is also an interaction energy between units.

Even if we knew nothing more about statistical mechanics and the Hopfield neural network than what we've read in this so far, we'd now know that the Hopfield neural network lives smack in the middle of the statistical mechanics universe, and that it is trained using a (free) energy minimization method. We'd also know that this neural network, even if not popular today (due to memory constraints for storing patterns in this network), is part-and-parcel of the world of neural networks and machine learning methods that use energy minimization.

We've given the most minimal and superficial attention to the notion of free energy. This concept has been foundational to modern neural networks (e.g., the Hopfield network), and continues with even broader scope and implications today, as it plays a key role in multiple machine learning methods.

Bibliography

[1] K. Friston, "Life as we know it," *Journal of The Royal Society Interface*, vol. 10, no. 86, 2013.

[2] J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 79, pp. 2554–2558, April 1982.